

## Molecular cloning of a bovine cathepsin

Nicholas J. GAY and John E. WALKER\*

*M.R.C. Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.*

(Received 9 August 1984/Accepted 12 October 1984)

A cDNA clone for a thiol endoproteinase has been isolated from a bovine heart cDNA library by using a mixture of 32 synthetic oligonucleotides as a hybridization probe. The inserted region is 672 base pairs in length. It contains a sequence encoding the C-terminal region of a protein that is homologous to rat liver cathepsins B and H and to plant thiol proteinases. In addition, it contains the sequence of 442 bases corresponding to the 3' untranslated region of the mRNA. The inserted region was used as a specific probe in RNA transfer analysis; the size of the mRNA encoding the thiol endoproteinase is estimated to be approx. 1.7 kilobases. Thus, the maximum size of the encoded protein is about 350–400 amino acids.

Neurosecretory peptides and polypeptide hormones are synthesized as larger precursors that undergo proteolytic processing during intracellular transport, storage and secretion (Docherty & Steiner, 1982). The signal sequence is first removed in the rough endoplasmic reticulum (Blobel & Dobberstein, 1975). Then further processing of the polypeptides that are formed takes place in the Golgi over a period of up to several hours, during which secretory vesicles mature to storage granules. Cleavage of proproteins often takes place at pairs of basic amino acids and may be followed by removal of the C-terminal basic amino acid thus generated, by a carboxypeptidase B-like enzyme (Loh *et al.*, 1984; Mains *et al.*, 1983; Steiner *et al.*, 1984). The enzyme involved in the first cleavage step is a thiol endoproteinase with a trypsin-like specificity (Docherty *et al.*, 1983). Thiol endoproteinases with a range of specificities occur in lysosomes (Barrett, 1977). Cathepsins B, H and L are among the most abundant, and cathepsins T, N, P and S appear to be minor constituents. Sequence analysis has demonstrated that cathepsins B and H are related to each other and to the plant thiol proteases papain and actinidin (Takio *et al.*, 1983). Cathepsin B and related proteinases have also been detected in insulin secretory granules by covalent affinity labelling (Docherty *et al.*, 1984). The mature protein ( $M_r$  31 500) has a broad specificity (McKay *et al.*, 1983), but it has

been suggested that a larger form, also detected in the secretory granules, may have a more restricted specificity favouring sequences of pairs of basic amino acids such as serve as processing signals (Docherty *et al.*, 1984). Thus, the molecular characterization of these precursor forms of cathepsins is of interest.

We have fortuitously isolated a cDNA clone from a cDNA library made from bovine heart mRNA that we have shown to contain a sequence encoding part of a protein related to the C-terminal region of cathepsins B and H. By hybridization with this clone we have detected a specific mRNA species in bovine mRNA estimated to be 1.7 kb in length. This mRNA could encode a protein of  $M_r$  40 000. Although this protein is probably not cathepsin B, the availability of the clone now makes it possible to characterize this cathepsin gene fully, and, given the close sequence relationship between cathepsin B and H, could provide the means of characterizing other members of this family also.

### Materials and methods

#### *Chemicals and reagents*

Nucleotide monomers were purchased from Cruachan Chemicals, Livingston, Scotland, U.K.; the Klenow fragment of DNA polymerase was from Boehringer, Lewes, Sussex, U.K. and avian myeloblastosis virus reverse transcriptase was from Anglian Biotechnology, Colchester, Essex, U.K.

Abbreviations used: kb, kilobases; bp, base pairs.

\* To whom correspondence and reprint requests should be sent.

### Oligonucleotide synthesis

A mixed oligonucleotide with the sequence



corresponding to the protein sequence Met-Ala-Ile-Glu-Glu-Gln, was synthesized by a solid phase phosphotriester method (Gait *et al.*, 1982).

### cDNA synthesis and cloning

Whole RNA was prepared from bovine heart tissue according to Chirgwin *et al.* (1979), except that the RNA was pelleted from the guanidinium thiocyanate homogenate through a cushion of 5.7M-CsCl/0.1M-EDTA, pH7.5, at 35000g for 16h. Poly(A)-containing RNA was selected on a column of oligo(dT)-cellulose by the method of Aviv & Leder (1972). First strand cDNA was synthesized from 20µg of poly(A)-containing RNA in a solution of 0.14M-KCl/0.1M-MgCl<sub>2</sub>/0.1M-Tris/HCl, pH8.3, containing 200µg of oligo(dT)<sub>12-18</sub>/ml, 200µg of mRNA/ml and 200 units of avian myeloblastosis virus reverse transcriptase/ml. This solution was incubated at 37°C for 90 min. RNA was removed by alkaline hydrolysis (50mM-NaOH/20mM-EDTA, 60°C, 1h). The solution was extracted once with an equal volume of phenol/chloroform (1:1, v/v) and passed over a 1 ml column of Sephadex G-50 (coarse grade) in TE buffer (10mM-Tris/HCl/10.1mM-EDTA, pH8.0). DNA was precipitated with ethanol and recovered by centrifugation.

Second strand cDNA was synthesized in a solution of 0.1M-Hepes (pH6.9)/10mM-MgCl<sub>2</sub>/2mM-dithiothreitol/0.07M-KCl containing 0.5mM each of the four deoxyribonucleotides and 500 units of the Klenow fragment of DNA polymerase/ml for 18h at 15°C. Double stranded cDNA was treated with S<sub>1</sub> nuclease (Vogt, 1973) (20 units/µg of first strand RNA) and then repaired with Klenow polymerase (Seeburg *et al.*, 1977). The cDNA was fractionated in a 1.5% HGT agarose gel and collected by trough electroelution to eliminate strands less than 300bp in length (Bankier & Barrell, 1983). The cDNA was then ligated with pUC8 plasmid DNA (Messing & Vieira, 1982) which had been digested with *Sma*I, and treated with calf intestinal phosphatase (Morton, 1955). Ligations were used to transform *Escherichia coli* TG1, an *Eco*K restriction-negative derivative of *E. coli* JM101 (Messing, 1979), by the method of Hanahan (1983). About 300000 recombinants were recovered and these were amplified by growth for 4.5h in SOB medium (Hanahan, 1983) containing 50µg of ampicillin/ml. The cells were recovered by centrifugation (1000g, 10min),

and resuspended in SOB medium containing 20% (v/v) glycerol and stored at -70°C.

### Screening of the cDNA library with oligonucleotides

Recombinants were grown on ampicillin plates at 37°C for 10h and transferred to sterile Whatman 541 paper filters. The filters were processed for hybridization as described by Gergen *et al.* (1979). The oligonucleotide probes were labelled using polynucleotide kinase and [ $\gamma$ -<sup>32</sup>P]ATP (specific radioactivity approx. 3000Ci/mmol) and purified by electrophoresis through a 20% polyacrylamide gel containing 6M-urea and 90mM-Tris (pH8.3)/90mM-boric acid/25mM-EDTA. After elution (Maxam & Gilbert, 1980) labelled oligonucleotides were further purified by centrifugation through a Sephadex G-50 column (1ml). Filters were hybridized with the oligonucleotides under the conditions described by Carroll & Porter (1983) and Bentley & Porter (1984), except that hybridization was at 38°C and the filters were washed at 45°C.

### Isolation and analysis of hybridizing recombinants

Colonies containing plasmids that hybridized with the probe were purified and plasmid DNA was prepared from 1.5ml of stationary phase cultures by the alkaline sodium dodecyl sulphate procedure (Birnboim & Doly, 1979). The recombinant DNA was released by digestion of the plasmid with *Pst*I and *Eco*RI or with *Bam*H1 and *Eco*RI and purified after electrophoresis through a 1% LGT agarose gel (Sanger *et al.*, 1980). The recombinant DNA was then ligated with M13 mp8 and M13 mp9 DNA that had been digested either with *Bam*H1 and *Eco*RI or with *Pst*I and *Eco*RI (Messing & Vieira, 1982). The insert contained two *Bam*H1 sites (see Fig. 1). DNA sequencing of the resultant M13 clones was carried out by the dideoxy chain termination method (Sanger *et al.*, 1977; Biggin *et al.*, 1983; Bankier & Barrell, 1983). Thus, the sequence was completely determined in both orientations of the DNA.

### 'Northern' blot analysis

Samples of poly(A)-containing and total RNA for 'Northern' blot analysis (Thomas, 1980) were reacted at 50°C for 10min in a solution containing 50% dimethyl sulphoxide, 10mM-NaH<sub>2</sub>PO<sub>4</sub>, pH6.9, and 0.13M-glyoxal. Samples were subject to electrophoresis in a 1.5% agarose gel in 10mM-NaH<sub>2</sub>PO<sub>4</sub>, pH6.9, and RNA was then transferred to nitrocellulose filters. The blot was hybridized with a single stranded 'prime cut' probe (Farrell *et al.*, 1983) containing the complementary sequence of the RNA. Hybridization was carried out for 18h at 65°C in a solution containing 4 × SSC (0.06M-trisodium citrate/0.6M-NaCl), 5 × Denhardt's solu-

S E Y N D Q A F I N H I V S V A G W G V  
 CTCGGAATACAACGACCAGGCCCTTCATAAACCACATCGTCTCCGTGGCCGGGTGGGGTGT  
 10 20 30 40 50 60

S D G M E Y W I V R N S W G E F W G E H  
 CAGCGATGGCATGGAGTACTGGATTGTCCGGAACCTCGTGGGGAGAACCATGGGGCGAGCA  
 70 80 90 100 110 120

G W M R I V T S T Y K G G E G A R Y N L  
 CGGCTGGATGCGGATCGTGACCAGCACCTACAAAGGTGGGGAGGGCGCCCGTTACAACCT  
 130 140 150 160 170 180

A I E E S C T F G D F I V \*  
 GGCCATCGAGGAGAGCTGCACGTTTGGGGACCCCATTTGTTTAAGGCAGAGCAGCTCTTAC  
 190 200 210 220 230 240

AGAAAGGATCGCGAGAGCCGGAACCAGAGGGGATCCCATTTGTCACAGGCACCGGGGTGGC  
 250 260 270 280 290 300

GCTGCCGTGGTTTGAAGGAACTGGGGGTTGACGTTACGTCACAGCCAGTGATGGCCCTG  
 310 320 330 340 350 360

AGCACCGAGGACAGGCACGGTGGAAAATGCCACACAGGCCTGACATGGGGGCGAGCCGGG  
 370 380 390 400 410 420

GAGCCGCGGGCTCCGCGCCTGCACTGGATGGCTTCCTGCCGGGAGAGCAGCCGGGAGAAG  
 430 440 450 460 470 480

CGGGATCCGAGGGGCGAGTGAACGATGTGACCTCCGTAGCAGATGACTTGGCAGCTGTGGA  
 490 500 510 520 530 540

CTGGGAGGAAAAACAGCTCGCACTCACCACCAGTTCCCTTTGTCACCTTGAAACCAATGGG  
 550 560 570 580 590 600

GCGCACAGGGGGGAGATGGTAATTTGAGTTGCCCAAGTGATGAATAAAATGCACACTTC  
 610 620 630 640 650 660

ACACCAAAAAA  
 670

Fig. 1. DNA sequence of inserted bovine cDNA in *pcCD1* and derived protein sequence

Two *Bam*HI sites that were useful in the sequence determination are underlined. The boxed sequence is likely to be the signal of addition of poly(A) to the 3' end of the message (Proudfoot & Brownlee, 1976).

tion (0.1% Ficoll/0.1% bovine serum albumin/0.1% polyvinylpyrrolidone 360), 100  $\mu$ g of sonicated salmon sperm DNA/ml and 0.5% n-lauryl sarcosine. Then the blot was washed in  $2 \times$  SSC at 65°C.

#### Computer analysis of DNA and protein sequences

The DNA sequences generated were compiled with the aid of the computer programs DBAUTO and DBUTIL (Staden, 1982a) and then analysed for various features with ANALYSEQ (Staden, 1984). Derived protein sequences were compared with the protein database of Doolittle (1981), using the rapid searching techniques of Wilbur & Lipman (1983) and subsequently, once matches had been detected, with DIAGON (Staden, 1982b).

## Results and discussion

#### Isolation and DNA sequence of cDNA clone

The cDNA library was constructed from bovine heart poly(A)-containing RNA and contained  $3 \times 10^5$  transformants. The average insert size was 300bp. The oligonucleotide probe had been designed as a probe for the  $\alpha$ -subunit of beef mitochondrial  $H^+$ -ATPase. This contains the protein sequence Met-Ala-Ile-Glu-Glu-Gln (V. L. J. Tybulewicz & J. E. Walker, unpublished work). Six positively hybridizing clones were identified with the probe.

#### Sequence of insert

The nucleotide sequences of the inserts were determined by the dideoxy chain termination method (Sanger *et al.*, 1977) as modified by Biggin *et al.* (1983). They were translated in all six possible reading frames using an option in ANALYSEQ (Staden, 1984). The derived protein sequences were screened against the database of Doolittle (1981) and against the protein sequence of the bovine  $H^+$ -ATPase  $\alpha$ -subunit (V. L. J. Tybulewicz & J. E. Walker, unpublished work).

One clone, p $\alpha$ CD1, was found to contain a sequence (Fig. 1) related to cathepsins B and H, papain and actinidin (Fig. 2). The remaining five clones are all derived from ferritin mRNA, the hybridizing region being in the 3'-untranslated region (N. J. Gay & J. E. Walker, unpublished work).

The relationships between the thiol endoproteinases, as shown in one case in Fig. 3, were verified by pairwise comparison with DIAGON. The alignment of the sequence of the p $\alpha$ CD1 insert with this family of thiol endoproteinases (Fig. 2) confirms that the clone codes for the C-terminal 73 residues of a protein belonging to this family. It is apparently most closely related to cathepsin B. Of particular significance is the histidine residue at position 9 in the bovine cathepsin. This residue is conserved throughout the thiol endoproteinases (see Fig. 2) and is the active site histidine in papain

	10	*	20	30		40	50	60	70
Bovine cDNA	S E Y N D Q A F I N	H	I V S V A	G W G V S D	G M E Y W I V R	N S W	G E		
Cathepsin B (189-252)	K H E A G D V M G G	H	A I R I L	G W G I E N	G V P Y W L V A	N S W	N V		
Cathepsin H (158-220)	S C H K T P D K V N	H	A V L A V	G Y G E Q N	G L L Y W I V K	N S W	G S		
Papain (151-212)	F V G P C G N K V D	H	A V A A V	G Y N P - - -	G Y I L I K	N S W	G T		
Actinidin (154-220)	F T G P C G T A V D	H	A I V I V	G Y G T E G	G V D Y W I V K	N S W	D T		
Bovine cDNA	P W G E H G	W M R	I V T S T - - -	Y K G	G E G A R Y N L A I E E S C T F G D P I V				
Cathepsin B	D W G D N G	F F K I	L R G E - - -	N H C	G I E S E I V A G I P R T Q				
Cathepsin H	N W G N N G	Y F L I	E R G K - - -	N M C	G L A A C A S Y P I P Q V				
Papain	G W G E N G	Y I R I	K R G T Q N S Y G V C	G L Y T S S F Y P V K N					
Actinidin	T W G E E G	Y M R I	L R N V G G A - G T C	G I A T M P S Y P V K Y N N					

Fig. 2. Homology between bovine cDNA and rat cathepsins B and H and plant thiol endoproteinases (see Takio *et al.*, 1983). For the sequences of papain see Light *et al.* (1964), Husain & Lowe (1969, 1970) and Mitchel *et al.* (1970) and for actinidin see Carne & Moore (1978). \* denotes the active site histidine in papain.

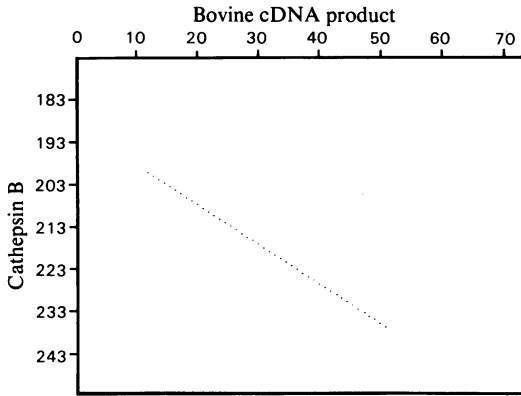


Fig. 3. Homology between rat cathepsin B and bovine cDNA product assessed with DIAGON

The calculation was made with a window size of 25 and a score of 280. The highest score reached was 359, corresponding to a double matching probability (McLachlan, 1971) of  $1.0 \times 10^{-11}$ .

(Drenth *et al.*, 1971). Another conserved sequence is Asn-Ser-Trp (residues 29–31 in bovine cathepsin). According to the structure of papain, this asparagine is hydrogen bonded to the active site histidine and tryptophan serves to stabilize the bond. The active site cysteine is found at position 25 in papain (Drenth *et al.*, 1971). The fragment of the bovine cathepsin presented here does not extend to this region.

The derived protein sequence contains the sequence Leu-Ala-Ile-Glu-Glu-Ser encoded by the DNA sequence CTG·GCC·ATC·GAG·GAG·AGC. This matches the mixed probe in 14 consecutive positions (underlined). Under the hybridization conditions employed, the duplex would have a melting temperature of 46°C.

The rest of the sequence is made up of the 3' untranslated region of the mRNA. Near base 650 is found the sequence AATAAA (boxed in Fig. 1). This corresponds to the RNA sequence UUAUUU, the signal for poly(A) addition (Proudfoot & Brownlee, 1976). This sequence is followed 17 bases later by A<sub>6</sub>, presumably the start of the 3' poly(A) tail itself.

#### Size of mRNA

This was determined by 'Northern' hybridization as shown in Fig. 4 and is estimated to be about 1.7 kb. Taking into account the 451 bases of 3' untranslated region this mRNA could potentially encode a protein with an  $M_r$  of 35 000–40 000, depending upon the size of the 5' untranslated region.

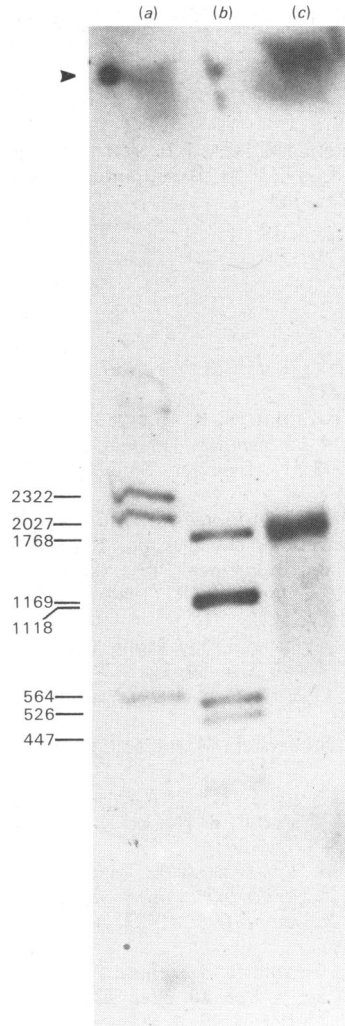


Fig. 4. 'Northern' hybridisations with bovine cDNA. The size of fragments resulting from *Hind*III digestion of (a) bacteriophage  $\lambda$  DNA and (b) SV40 DNA are shown. (c) Poly(A)-containing heart RNA (approx. 5  $\mu$ g) probed with a  $^{32}$ P-labelled prime-cut DNA probe corresponding to the whole of the cloned region. Whole unselected heart RNA, hybridized with the same probe (results not shown) gave a weak band running in the same position as the band in (c). The arrow denotes the origin. For further details see the Materials and methods section.

Further information concerning the identity of the mRNA and the protein should be obtained from the complete sequence of the gene. However, at this point it seems unlikely that the cDNA codes for cathepsin B as the N-terminal sequences in the

rat and cow enzymes are very similar (Pohl *et al.*, 1982).

We are grateful to Drs. A. J. Barrett and G. S. Salvesen for discussions. N. J. G. is supported by an M.R.C. Training Fellowship.

## References

- Aviv, L. & Leder, P. (1972) *Proc. Natl. Acad. Sci. U.S.A.* **69**, 1408–1412
- Bankier, A. T. & Barrell, B. G. (1983) in *Techniques in Nucleic Acid Biochemistry* (Flavell, R. A., ed.), pp. B508/1–B508/31, Elsevier Scientific Publishers, Ireland
- Barrett, A. J. (1977) in *Proteases in Mammalian Cells and Tissues* (Barrett, A. J., ed.), pp. 181–208, Elsevier/North Holland Biomedical Press, Amsterdam
- Bentley, D. R. & Porter, R. R. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1212–1215
- Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3963–3965
- Birnboim, H. C. & Doly, T. (1979) *Nucleic Acids Res.* **7**, 1513–1523
- Blobel, G. & Dobberstein, B. (1975) *J. Cell Biol.* **67**, 835–851
- Carne, A. & Moore, C. (1978) *Biochem. J.* **173**, 73–83
- Carroll, M. C. & Porter, R. R. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 264–267
- Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299
- Docherty, K. & Steiner, D. F. (1982) *Annu. Rev. Physiol.* **44**, 625–638
- Docherty, K., Carroll, R. & Steiner, D. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3245–3249
- Docherty, K., Hutton, J. C. & Steiner, D. F. (1984) *J. Biol. Chem.* **259**, 6041–6044
- Doolittle, R. F. (1981) *Science* **214**, 149–159
- Drenth, J., Jansonius, J. N., Koekoek, R. & Walthers, B. G. (1971) *Adv. Protein Chem.* **25**, 79–115
- Farrell, P. J., Deininger, P. L., Bankier, A. & Barrell, B. G. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1565–1569
- Gait, M. J., Matthes, H. W. D., Singh, M., Sproat, B. S. & Titmas, R. C. (1982) *Nucleic Acids Res.* **10**, 6243–6254
- Gergen, J. P., Stern, R. H. & Websink, P. C. (1979) *Nucleic Acids Res.* **7**, 2115–2136
- Hanahan, D. (1983) *J. Mol. Biol.* **166**, 557–580
- Husain, S. S. & Lowe, G. (1969) *Biochem. J.* **114**, 279–288
- Husain, S. S. & Lowe, G. (1970) *Biochem. J.* **116**, 689–692
- Light, A., Frater, R., Kimmel, J. R. & Smith, E. L. (1964) *Proc. Natl. Acad. Sci. U.S.A.* **52**, 1276–1283
- Loh, Y. P., Brownstein, M. J. & Gainer, H. (1984) *Annu. Rev. Neurosci.* **7**, 189–222
- Mains, R. E., Eipper, B. A., Glembotski, C. C. & Doves, R. M. (1983) *Trends Neurochem. Sci.* **6**, 229–235
- Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560
- McLachlan, A. D. (1971) *J. Mol. Biol.* **61**, 409–424
- McKay, M. J., Offermann, M. K., Barrett, A. J. & Bond, J. S. (1983) *Biochem. J.* **213**, 467–471
- Messing, J. (1979) *Recombinant DNA Technical Bull.* **2**, 43–48
- Messing, J. & Vieira, J. (1982) *Gene* **19**, 269–276
- Mitchel, R. E. J., Chaiken, I. M. & Smith, E. L. (1970) *J. Biol. Chem.* **245**, 3485–3492
- Morton, R. K. (1955) *Biochem. J.* **60**, 573–577
- Pohl, J., Baudys, M., Tomasek, V. & Kostka, V. (1982) *FEBS Lett.* **142**, 23
- Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263**, 211–214
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5476
- Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143**, 161–178
- Seeburg, P. H., Shine, J., Martial, J. A., Baxter, J. D. & Goodman, H. M. (1977) *Nature (London)* **220**, 486–494
- Staden, R. (1982a) *Nucleic Acids Res.* **10**, 4731–4751
- Staden, R. (1982b) *Nucleic Acids Res.* **10**, 2951–2961
- Staden, R. (1984) *Nucleic Acids Res.* **12**, 521–538
- Steiner, D. F., Docherty, K. & Carroll, R. (1984) *J. Cell Biochem.* **24**, 121–130
- Takio, K., Towatari, T., Katunuma, N., Teller, D. C. & Titani, K. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3666–3670
- Thomas, P. S. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 5201–5204
- Vogt, V. M. (1973) *Eur. J. Biochem.* **33**, 192–195
- Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 726–730